CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting *in vivo*

Miguel A Moreno-Mateos[1,4], Charles E Vejnar[1,4], Jean-Denis Beaudoin[1], Juan P Fernandez[1], Emily K Mis[1,2], Mustafa K Khokha[1,2] & Antonio J Giraldez[1,3]

CRISPR-Cas9 technology provides a powerful system for genome engineering. However, variable activity across different single guide RNAs (sgRNAs) remains a significant limitation. We analyzed the molecular features that influence sgRNA stability, activity and loading into Cas9 *in vivo*. We observed that guanine enrichment and adenine depletion increased sgRNA stability and activity, whereas differential sgRNA loading, nucleosome positioning and Cas9 off-target binding were not major determinants. We also identified sgRNAs truncated by one or two nucleotides and containing 5′ mismatches as efficient alternatives to canonical sgRNAs. On the basis of these results, we created a predictive sgRNA-scoring algorithm, CRISPRscan, that effectively captures the sequence features affecting the activity of CRISPR-Cas9 *in vivo*. Finally, we show that targeting Cas9 to the germ line using a Cas9-nanos 3′ UTR led to the generation of maternal-zygotic mutants, as well as increased viability and decreased somatic mutations. These results identify determinants that influence Cas9 activity and provide a framework for the design of highly efficient sgRNAs for genome targeting *in vivo*.

Genome-editing systems are essential for understanding gene function by means of reverse genetics. Zinc finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs) have been broadly used to generate short insertion or deletion (indel) mutations[1,2]. These techniques have enabled genetic engineering in multiple biological systems through protein-based recognition of DNA, but they are limited by their variable efficiency and cumbersome assembly. The prokaryotic CRISPR-Cas9 machinery[3] was recently adapted to eukaryotic cells through use of the endonuclease Cas9 and an sgRNA with complementarity to the target DNA[4–6]. It has been successfully used to induce targeted genetic mutations in several model organisms such as *Caenorhabditis elegans*, *Drosophila*, mouse and zebrafish[7–10]. However, the variable activity of different sgRNAs still results in inconsistent CRISPR-Cas9 activity. The molecular features that determine sgRNA stability, loading and targeting *in vivo* remain largely unexplored. Recently, studies in human and mouse cell lines have identified features shown to modulate CRISPR-Cas9 activity, as assessed through phenotypic selection[11,12]. Thus, the effects of extrinsic features such as the microenvironment mediating double-strand-break DNA repair and selection for deleterious mutations are difficult to separate from the effects of intrinsic features of the sgRNA. Indeed, one of the strongest features correlating with high sgRNA activity is a depletion of uridines in the sgRNA, which in fact is related to the termination signal of the Pol III transcription machinery used to express the sgRNAs[11]. Circumventing this problem, Gagnon and collaborators targeted more than 100 genes in zebrafish embryos, each with a single sgRNA[13]. They confirmed that a guanine adjacent to the protospacer-adjacent motif (PAM) favors cleavage, in agreement with studies in cell lines[11,12], and suggested other sequence features that correlate with targeting efficiency. However, it is unclear how much these features are influenced by the specific genomic loci tested, given that only one sgRNA was evaluated per gene. This could limit the capacity to identify specific features that mediate sgRNA activity, such as stability, Cas9 loading and target recognition. Consequently, the principles underlying effective targeting using the CRISPR-Cas9 system *in vivo* are still largely unknown.

In this study, we analyzed the stability, loading and mutagenic activity of 1,280 sgRNAs targeting 128 genes. We found that efficient sgRNAs have a biased sequence composition that affects CRISPR-Cas9 stability and activity. We compared the efficiency of 640 alternative sgRNAs (truncated, extended and 5′ mismatch containing) increasing the number of target sites in the genome. We integrated these findings into a predictive model, CRISPRscan (http://CRISPRscan.org). Finally, we designed a Cas9 construct that targets mutations preferentially in the germ line, reducing the number of somatic mutations that allow the generation of maternal mutants. Together these findings improve the CRISPR-Cas9 system by providing insights into the factors that influence sgRNA activity.

RESULTS

Measuring the activity of >1,000 sgRNAs

To determine which factors influence the generation of sgRNA-mediated DNA lesions, we first measured the activity of 1,280 sgRNAs targeting 128 different genes in the zebrafish genome. For each gene, we designed a total of ten sgRNAs falling within a 1.2-kb span (**Supplementary Fig. 1**), with the majority (96%) targeting exons. We reasoned that *in vitro* transcribed sgRNAs

[1]Department of Genetics, Yale University School of Medicine, New Haven, Connecticut, USA. [2]Department of Pediatrics, Yale University School of Medicine, New Haven, Connecticut, USA. [3]Yale Stem Cell Center, Yale University School of Medicine, New Haven, Connecticut, USA. [4]These authors contributed equally to this work. Correspondence should be addressed to A.J.G. (antonio.giraldez@yale.edu).

RECEIVED 20 APRIL; ACCEPTED 12 JULY; PUBLISHED ONLINE 31 AUGUST 2015; DOI:10.1038/NMETH.3543

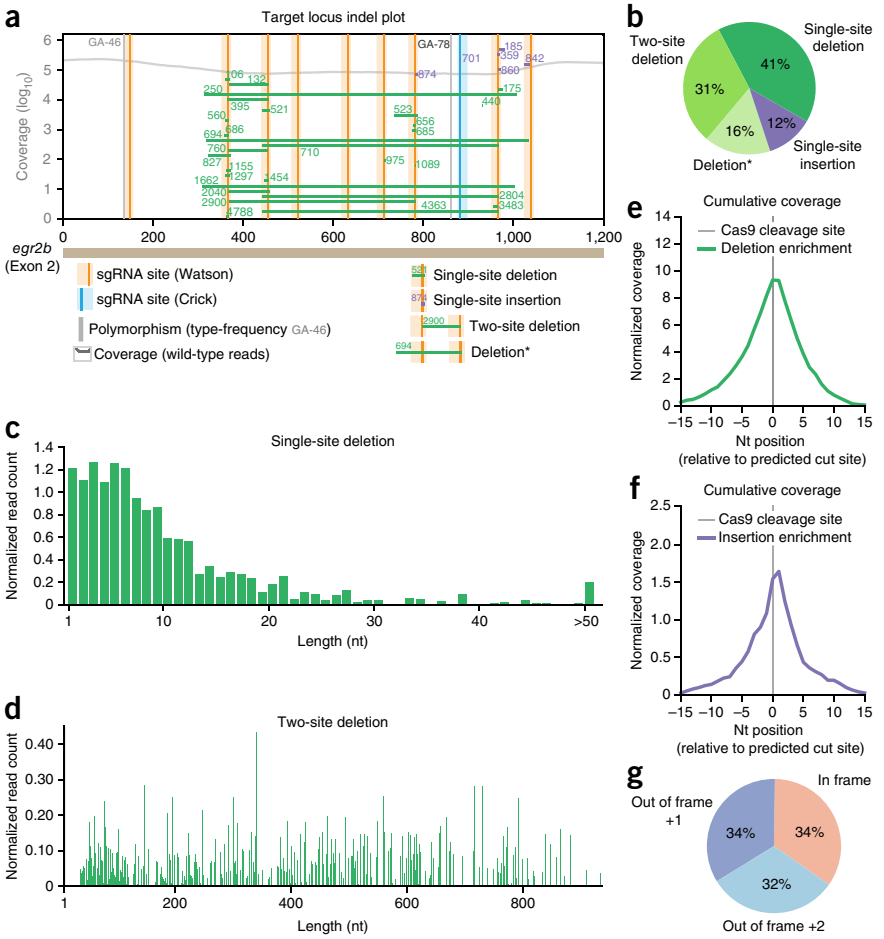

Figure 1 | Measuring the activity of >1,000 sgRNAs. (**a**) Deletions and insertions found on the *egr2b* gene locus with read coverage (gray curve). Wider vertical bars represent sgRNA target sites, and the thin solid lines indicate the Cas9 cleavage site. Horizontal bars represent deletions (green) and insertions (purple), with the supporting number of reads shown next to each bar. Vertical gray bars show polymorphisms. *Deletion with boundaries that cannot be unambiguously assigned to a single-site or two-site deletion (Online Methods). (**b**) Distribution of sgRNA mutations caused by a single site or between two sites. (**c**) Lengths of deletions induced by single sgRNAs (median of 7 nt). (**d**) Lengths of deletions induced by sgRNA pairs (median of 400 nt). (**e**,**f**) Cumulative coverage of mutated reads overlapping single-site deletions (**e**) and insertions (**f**) normalized by wild-type reads. (**g**) Distribution of frame shifts caused by all CRISPR-Cas9–induced mutations.

(**Supplementary Fig. 1e**,**g**) would allow us to measure the influence of sgRNA stability, Cas9 loading and target-sequence composition independently of transcription rates that might differ in DNA-based libraries typically used in cell culture. To this end, we first injected 16 pools of 80 sgRNAs (targeting eight loci per pool) together with Cas9-encoding mRNA into zebrafish embryos at the one-cell stage (**Supplementary Fig. 1b**). We analyzed indel frequency at 9 h post-fertilization (hpf) using high-throughput sequencing, and we also collected embryos at 0 and 1.25 hpf to measure input sgRNA levels (**Supplementary Fig. 1c**,**d**, **Supplementary Table 1**, **Supplementary Data Set 1** and Online Methods). As a control, we sequenced uninjected sibling embryos and discarded 37 target sites presenting polymorphism with the reference zebrafish genome. We measured individual sgRNA activity as the number of indel-containing reads divided by the total number of reads in each site. We observed that the vast majority of sgRNAs induced deletions (88%) (**Fig. 1a**,**b** and **Supplementary Data Set 1**). Their activity ranged from 0% to 77%, even among sgRNAs in the same locus, indicating that the wide range of efficiency depends on additional factors that are independent of the targeted locus. Deletions caused by individual sgRNAs had a median length of 7 nt and represented >40% of the events (**Fig. 1b**,**c**). Insertions were shorter (median of 4 nt) and represented 12% of the events (**Fig. 1b** and **Supplementary Fig. 1f**). Because multiple different sgRNAs targeting the same locus were coinjected, we observed a large fraction of deletions spanning two sgRNA target sites (two-site deletion) (31%), with a median length of 400 nt, although some were as many as 900 nt long (**Fig. 1b**,**d**). Indels had the maximum enrichment at the predicted position of the cleavage site, 3 nt upstream of the PAM motif[14], and showed no 5′ or 3′ bias (**Fig. 1e**,**f** and **Supplementary Fig. 1g**). Indeed, when we calculated the percentage of indels that maintained the frame or caused frame shifts, we found an even distribution of frames after repair (**Fig. 1g**). Taken together, these results indicate that sgRNA-induced mutations are associated with a wide range of efficiencies and cause mainly deletions. This provides a valuable data set for identifying the determinants that influence cleavage efficiency and for deriving better rules for predicting CRISPR-Cas9 activity.

Stable sgRNAs are more active, G rich and A depleted

DNA-mediated delivery of sgRNAs typically involves constitutive expression systems. In contrast, direct delivery of *in vitro* transcribed sgRNAs makes it possible to isolate the effects of sgRNA stability and loading into the Cas9 complex. To this end, we compared sgRNA levels at 0 hpf (injected input) and 1.25 hpf using high-throughput sequencing (**Supplementary Figs. 1b** and **2** and Online Methods). Analysis of the sgRNA nucleotide composition showed reproducible and significant enrichment of guanine and depletion of adenine in the more stable sgRNAs ($P = 3.7 \times 10^{-65}$ for guanine enrichment and $P = 1.5 \times 10^{-8}$ for adenine depletion; **Fig. 2a**,**b** and **Supplementary Fig. 2c**). Next, we analyzed whether stability could modulate sgRNA activity. We partitioned the sgRNAs into stable and unstable groups by comparing levels in the input to those at 1.25 hpf (**Fig. 2c** and **Supplementary Fig. 2d**). Although these two groups had similar levels at 0 hpf, the mutagenic activity of the stable sgRNAs was significantly higher than that of the unstable sgRNAs ($P = 6.9 \times 10^{-12}$, Mann-Whitney *U*-test) (**Fig. 2d** and **Supplementary Fig. 2e**). Additionally, mutagenic activity and sgRNA levels were significantly correlated ($r = 0.39$, $P = 2.1 \times 10^{-39}$; **Supplementary Fig. 3a**). Thus, sgRNA stability represents an important determinant of sgRNA function.

Guanine- and cytosine-rich RNAs fold into stable structures. Thus, we hypothesized that sgRNA folding energy could explain

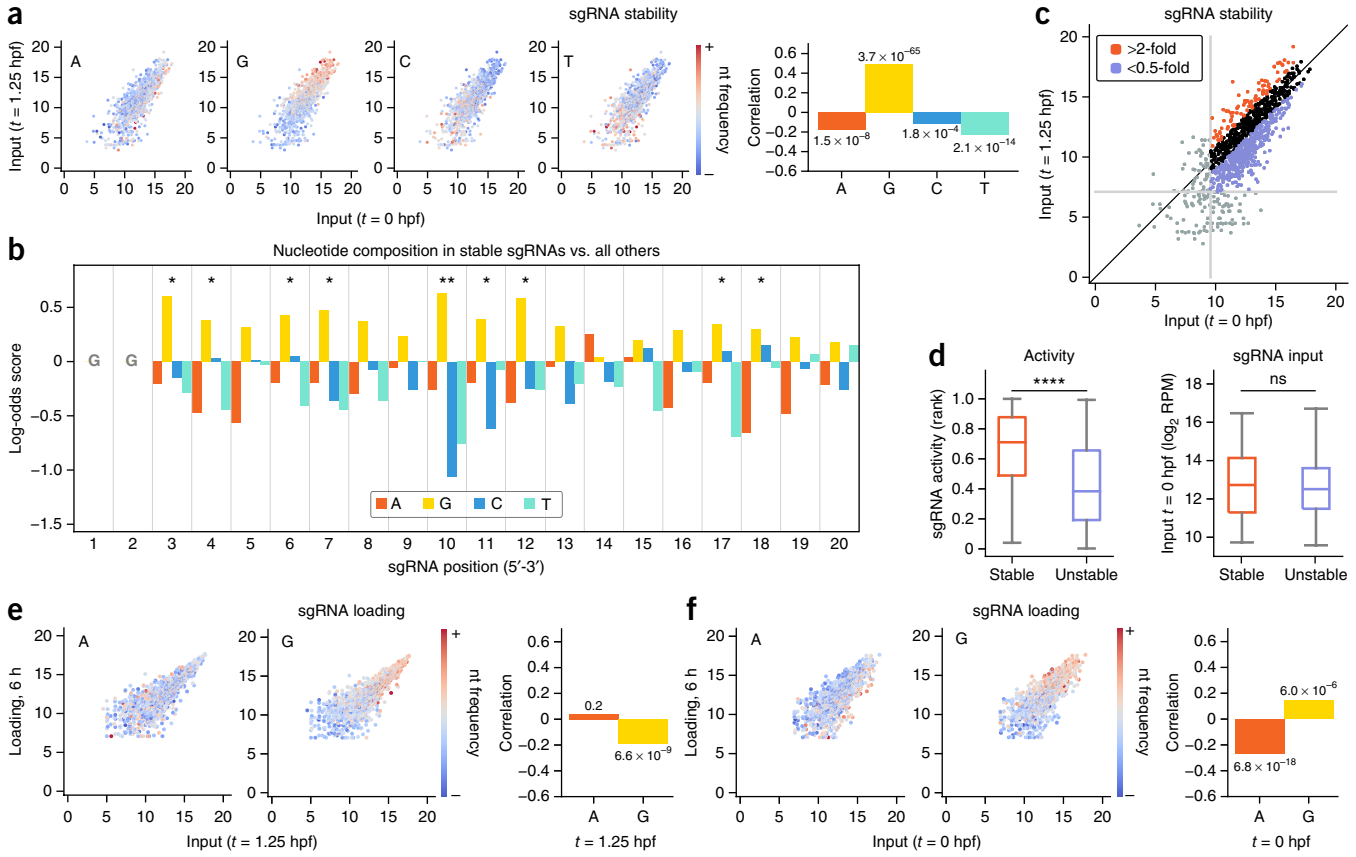

Figure 2 | Stable sgRNAs are more active, G rich and A depleted. (**a**) Biplot comparing sgRNA levels (log_2 reads per million mapped reads (RPM)) at 0 and 1.25 hpf, colored according to the key to indicate the frequencies of the 4 nt in each sgRNA. Corresponding Spearman correlations between nucleotide frequencies and sgRNA stability (ratio of levels at 1.25 hpf to those at 0 hpf) are shown on the right, with *P* values indicated within the graph. (**b**) The nucleotide composition of the most stable 20% of sgRNAs compared with the compositions of all other sgRNAs. Bars show log-odds scores of nucleotide frequencies for each position in the sgRNA (1–20) (*G*-test: *<0.05, **<0.01). (**c**) Biplot illustrating stable and unstable groups of sgRNAs, defined by greater than twofold enrichment or depletion between 0 and 1.25 hpf (log_2 RPM). sgRNAs with low read counts (bottom 10%) were excluded (gray dots). (**d**) Box-and-whisker plots (box spans first to last quartiles; whiskers represent 1.5× the interquartile range) showing sgRNA activity (left) and the input levels (right) in the stable and unstable sgRNAs. ****$P < 0.0001$, Mann-Whitney *U*-test; ns, not significant. (**e**,**f**) Biplots comparing sgRNA levels (log_2 RPM) loaded into Cas9 at 6 hpf to those at 1.25 hpf (**e**) and 0 hpf (**f**), colored to indicate the frequencies of A and G in each sgRNA. Corresponding Spearman correlations between nucleotide frequencies and sgRNA stability (ratio of levels loaded at 6 h to those at 1.25 hpf) are shown (right), with *P* values indicated in the graph.

the guanine enrichment in stable sgRNAs. To test this, we folded *in silico* all sgRNAs (including the 80-nt tail) to compute their ensemble free energies (EFEs). We observed a significant anticorrelation between sgRNA stability and EFE ($r = -0.495$, $P = 2.2 \times 10^{-62}$) (**Supplementary Fig. 3b**). sgRNA EFEs were also significantly anticorrelated with guanine content ($r = -0.475$, $P = 3.8 \times 10^{-57}$), but they were weakly correlated with cytosine content ($r = 0.073$, $P = 0.0021$), which suggests that guanine enrichment is not a mere consequence of lower folding energy. We postulated that guanines may protect against 5′-directed exonuclease degradation, as this enrichment occurred more prominently at the 5′ end of the sgRNA (**Fig. 2b** and **Supplementary Fig. 2c**). Guanine-rich sequences can fold into stable noncanonical structures called G-quadruplexes *in vivo*[15]. Indeed, a stronger correlation between guanine content and stability was observed when the sgRNA contained more than eight guanines, which is the minimal requirement to form a G-quadruplex (**Supplementary Fig. 3c**). To further support this, we tested a subset of stable and unstable sgRNAs for their ability to fold into G-quadruplex structures using in-line probing[16]. Notably, seven out of nine guanine-rich stable sgRNAs tested were able to form G-quadruplexes, compared with none of the five unstable sgRNAs (**Supplementary Fig. 3d**,**e** and **Supplementary Table 2**). These results suggest that G-quadruplex formation contributes to sgRNA stability.

Next, we used Flag-Cas9 to examine whether the loading of sgRNAs into Cas9 is influenced by the nucleotide sequence (**Supplementary Figs. 2a**,**b** and **3f**). We compared the total sgRNA composition at 1.25 hpf with the composition of the sgRNAs loaded into Flag-Cas9 after immunoprecipitation at 6 hpf. We observed a small but significant guanine depletion, which suggested that Cas9 might slightly disfavor loading of guanine-rich sgRNAs ($P = 6.6 \times 10^{-9}$; **Fig. 2e** and **Supplementary Fig. 2f**,**g**). However, this effect was negligible compared to the guanine enrichment observed between 0 and 1.25 hpf (**Fig. 2a**,**b** and **Supplementary Fig. 2c**). Indeed, when we compared loaded sgRNA levels to the input at 0 hpf, we still observed significant guanine enrichment, together with adenine depletion ($P = 6.0 \times 10^{-6}$; **Fig. 2f** and **Supplementary Fig. 2h**,**i**), correlated with a higher level of activity (**Supplementary Fig. 2j**,**k**). Preferentially loaded sgRNAs were also slightly enriched for cytosines in several positions (**Supplementary Fig. 2h**,**i**). Time-course analysis of sgRNAs loaded into Cas9 at 6 hpf and 9 hpf suggested that sgRNAs are stably retained in Cas9 once they are

loaded, as we did not detect a significant difference between time points (**Supplementary Fig. 3g**). Together, these results show that sgRNA stability *in vivo* is strongly influenced by the nucleotide composition, being favored by guanines and disfavored by adenines, and that this modulates sgRNA activity.

CRISPR-Cas9 activity is modulated by the sgRNA sequence

To test whether the mutagenic activity of sgRNAs is influenced by their sequence, we examined the nucleotide composition of efficient sgRNAs for each position in the target site extended by six nucleotides upstream and downstream. To exclude the effect of sgRNA stability, we normalized the activity of the sgRNA by its input level at 1.25 hpf. We identified ten positions with significantly different nucleotide distributions in the most efficient sgRNAs (top 20%) (**Fig. 3a**). First, nucleotides distal to the PAM (positions 1–14) were dominated by guanine enrichment, whereas positions 15–18 were characterized by cytosine enrichment. Thymidine and adenine nucleotides were depleted overall, with the exception of positions 9 and 10. Second, we observed strong guanine enrichment at position 20 of the sgRNA (**Fig. 3a**), consistent with previous studies[11,13]. Completing this observation, we found a corresponding depletion in cytosine at position 20 *in vivo*[11]. Furthermore, we observed that (i) guanine and cytosine were enriched at the first nucleotide of the PAM sequence, reconciling independent observations by Kuscu *et al.*[17] and Doench *et al.*[11], and (ii) position 3 was strongly enriched for guanine and depleted for cytosine. Lastly, outside the sgRNA-binding site we observed a guanine depletion one nucleotide downstream of the PAM. Thus, specific sgRNA sequences mediate efficient sgRNA-Cas9 target recognition and cleavage.

Next, we integrated these observations into a model to predict mutagenic activity on the basis of sgRNA target sequences. We used randomized logistic regression to select stable features that were the strongest determinants of sgRNA efficiency. Using these selected features, we trained a linear regression model on ranked sgRNA activity normalized by the input at 0 hpf (**Supplementary Table 3**). The resulting sgRNA-scoring method, which we named CRISPRscan, performed strongly when we compared its output to our experimental data ($r = 0.58$, $P = 7.1 \times 10^{-93}$, Pearson's correlation between predicted and experimental sgRNA activity). For example, of the top-scoring quintile of sgRNAs (CRISPRscan score > 0.6), 54% were among the sgRNAs determined to be most active by experimental evaluation, and only 2% were among the least active (**Fig. 3b**). To ensure that our prediction model was not

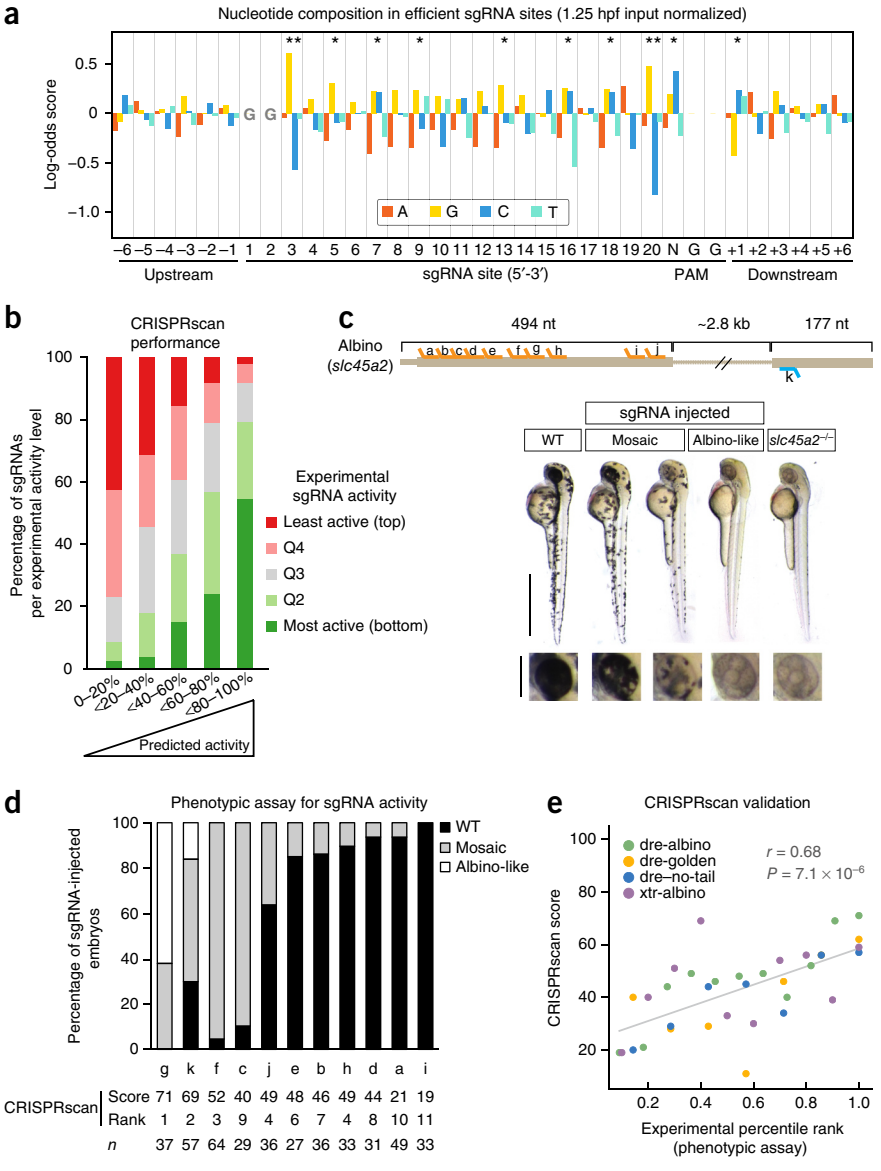

Figure 3 | CRISPR-Cas9 activity is modulated by the sgRNA sequence. (**a**) The nucleotide composition of the most efficient 20% of sgRNA sites (positions 1–20 extended by the PAM sequence and 6 nt) compared with the composition of all other sites. Bars show log-odds scores of nucleotide frequencies for each position (*G*-test: *<0.05, **<0.01). (**b**) Performance of linear regression–based prediction model (CRISPRscan). sgRNAs were divided into quintiles on the basis of CRISPRscan scores (horizontal axis), and then each quintile was evaluated on the basis of its experimentally determined activity (color-coded as in key). (**c**) Diagram showing 11 sgRNA sites targeting *slc45a2* exons 1 and 2 used in an independent validation of the prediction model. Lateral views and insets of the eyes of 48-hpf embryos (bottom) show phenotypes obtained after sgRNA injection, demonstrating different levels of mosaicism compared with the wild type (WT). The rightmost picture is of an *slc45a2* loss-of-function mutant (−/−) described by White *et al.*[19]. Scale bar, 1 mm (0.25 mm in insets). (**d**) Phenotypic evaluation of 11 sgRNAs targeting *slc45a2*. Stacked bar plots show the percentage of albino-like (white), mosaic (gray) and phenotypically wild-type (black) embryos 48 hpf after injection. Predicted CRISPRscan scores, ranks and number of embryos evaluated (*n*) are shown for each sgRNA. (**e**) Correlation between CRISPRscan score and experimentally measured activity on the basis of all phenotypes used to independently validate CRISPRscan (**c**,**d** and **Supplementary Fig. 4**). The Spearman correlation and *P* value are indicated.

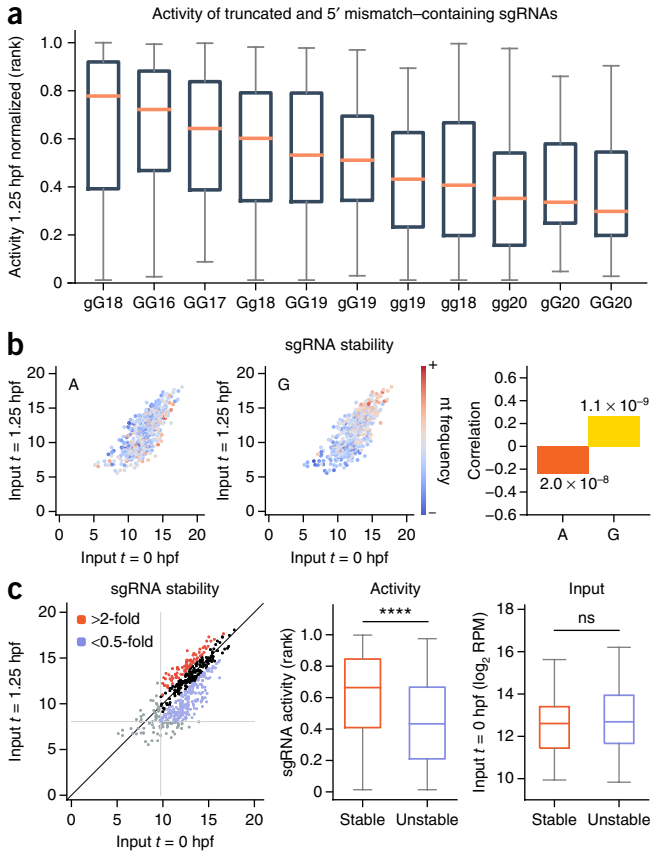

Figure 4 | Extending the CRISPR target repertoire with truncated, extended and 5′ mismatch–containing sgRNAs. (**a**) Ranked normalized activity of each class of alternative sgRNAs, ordered by median activity. Shorter sgRNAs (GG16 and GG17) and sgRNAs inducing one mismatch in the 5′ GG (gG18 and Gg18) were the most active alternatives to the canonical GG18 sgRNA. (**b**) Biplot comparing sgRNA levels (log_2 RPM) at 0 and 1.25 hpf, colored as in key to indicate the frequencies of A and G in each sgRNA. Corresponding Spearman correlations between nucleotide frequencies and sgRNA stability (ratio of levels at 1.25 hpf to those at 0 hpf) are shown at right, with *P* values indicated in the graph. (**c**) Biplot illustrating stable and unstable groups of sgRNAs, defined as those with greater than twofold enrichment or depletion between 0 and 1.25 hpf (log_2 RPM) (left). sgRNAs with low read counts (bottom 10%) were excluded (gray dots). Box-and-whisker plots show sgRNA activity and input levels in the stable and unstable sgRNAs. In **a** and **c**, boxes span first to last quartiles, and whiskers represent 1.5× the interquartile range. ****$P < 0.0001$, Mann-Whitney *U*-test.

overtrained and was generalizable to other sgRNAs, we performed cross-validations (Online Methods), which confirmed the performance of our model ($r = 0.45$, s.d. = 0.071). We observed that efficient sgRNA had a guanine enrichment at position 20 (**Fig. 3a**), which was the second highest feature selected by CRISPRscan within a dinucleotide $G_{19}G_{20}$ motif[18] (**Supplementary Table 3**). Finally, we independently validated CRISPRscan by determining the efficiency of 35 different sgRNAs targeting the albino (*slc45a2*), golden (*slc24a5*) and no-tail (*ntla*) loci in zebrafish, as well as the *slc45a2* locus in *Xenopus tropicalis* (**Fig. 3c**,**d** and **Supplementary Fig. 4**) (ref. 19). We observed a significant high correlation between CRISPRscan scores and phenotypic experimental activities ($r = 0.68$, $P = 7.1 \times 10^{-6}$) (**Fig. 3e**). In most cases at least two sgRNAs with the highest CRISPRscan scores were the most active *in vivo*. On the basis of the experimental validation, efficient sgRNAs scored above 0.55, and highly efficient ones scored above 0.70. We also tested whether the mutagenic activity of sgRNA-Cas9 complexes was influenced by the accessibility of the chromatin at the targeted locus or competition by putative off-target binding sites, but these did not significantly affect CRISPR-Cas9 activity (**Supplementary Fig. 5**). Together, these results show that the sequence composition of the sgRNA and the target strongly influence the mutagenic activity of the CRISPR-Cas9 system and that CRISPRscan successfully identifies the most active sgRNAs that mediate mutagenesis *in vivo*.

Alternative sgRNA formulations have variable activity

The sequence specificity of the T7 or SP6 promoters restricts 5′ sgRNA sequences to GG or GA, limiting the number of available targets *in vivo*. To extend the repertoire of potential targeting sites in the genome, we evaluated sgRNAs of various lengths (18–22 nt) with up to two nucleotide mismatches in positions 1 and 2 of the sgRNA (mismatch denoted by "g"); these were termed alternative sgRNAs (**Supplementary Fig. 6**). We compared the activity of 640 alternative sgRNAs spanning 11 different sgRNA formulations targeting a total of 64 loci (**Supplementary Fig. 6a–c**, **Supplementary Table 1** and **Supplementary Data Set 2**). We observed significantly different activities among them ($P = 1.0 \times 10^{-9}$, Kruskal-Wallis *H*-test). The most efficient alternatives to the canonical sgRNAs were sgRNAs shorter at the 5′ end by 1–2 nt[20,21] or sgRNAs of canonical length but inducing one mismatch in the 5′ GG (**Fig. 4a**). In contrast, longer sgRNAs were less effective, particularly those containing a 22-nt binding sequence. The average activity of alternative sgRNAs decreased with sequence variants in the following order: gG18 ~ GG16 ~ GG17 ~ Gg18 > GG19 ~ gG19 > gg19 ~ gg18 > gg20 ~ gG20 > GG20 (**Fig. 4a** and **Supplementary Fig. 6b**). Consistent with the stability features described above for canonical sgRNAs, alternative sgRNAs were more stable when enriched in guanine and depleted in adenine, which resulted in higher activity (**Fig. 4b**,**c** and **Supplementary Fig. 6d**). Stability among the different types of alternative sgRNAs was not significantly different ($P = 0.99$, χ^2 test). Next, we directly compared the activity of the canonical and alternative sgRNAs targeting the same site in the albino and golden loci (GG17 and GG16 versus GG18; GA18 versus Gg18) (**Supplementary Fig. 6e–h**). Most sgRNAs (eight out of ten) showed no significant difference in the generation of biallelic mutation, as quantified on the basis of the loss of pigmentation in the F0 injected embryos (**Supplementary Fig. 6f**,**h**). Thus, the activity of the most efficient alternative sgRNAs was equivalent to that of canonical sgRNAs. Indeed, we were able to adapt CRISPRscan to predict the activity of Gg18, GG17 and gG18, but not of GG16 (**Supplementary Fig. 6i**,**j**). These results suggest that CRISPRscan can infer the efficiency of those sgRNAs that increase the number of potential target sites in the zebrafish genome by eightfold (from ~5×10^6 to ~44×10^6 sites).

Targeting CRISPR-Cas9 activity to germ cells

CRISPRscan efficiently detected the most active sgRNAs *in vivo*. However, biallelic mutations derived from the use of highly efficient sgRNAs can result in a lethal phenotype for many essential genes (**Fig. 5a** and **Supplementary Figs. 4c** and **7**). Concentrating the mutagenic activity of Cas9 in the germ line would (i) minimize lethality due to somatic mutations and (ii) allow biallelic mutations in the germ cells removing the maternal contribution of the

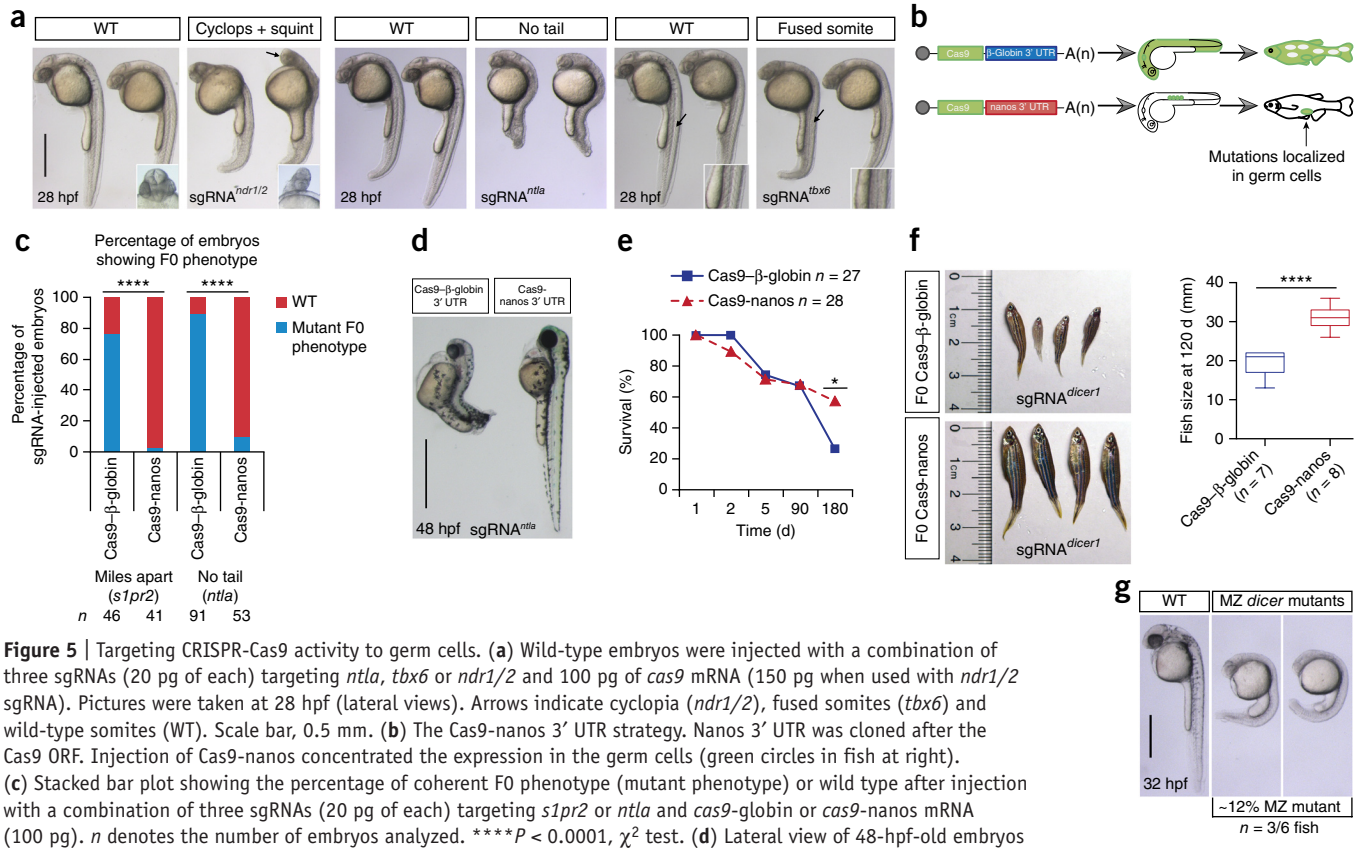

Figure 5 | Targeting CRISPR-Cas9 activity to germ cells. (**a**) Wild-type embryos were injected with a combination of three sgRNAs (20 pg of each) targeting *ntla*, *tbx6* or *ndr1/2* and 100 pg of *cas9* mRNA (150 pg when used with *ndr1/2* sgRNA). Pictures were taken at 28 hpf (lateral views). Arrows indicate cyclopia (*ndr1/2*), fused somites (*tbx6*) and wild-type somites (WT). Scale bar, 0.5 mm. (**b**) The Cas9-nanos 3′ UTR strategy. Nanos 3′ UTR was cloned after the Cas9 ORF. Injection of Cas9-nanos concentrated the expression in the germ cells (green circles in fish at right). (**c**) Stacked bar plot showing the percentage of coherent F0 phenotype (mutant phenotype) or wild type after injection with a combination of three sgRNAs (20 pg of each) targeting *s1pr2* or *ntla* and *cas9*-globin or *cas9*-nanos mRNA (100 pg). *n* denotes the number of embryos analyzed. ****$P < 0.0001$, χ^2 test. (**d**) Lateral view of 48-hpf-old embryos injected with the same sgRNA-Cas9 combinations described in **c** (targeting *ntla*). Scale bar, 1 mm. (**e**) Survival curves for fish injected with Cas9-nanos or Cas9–β-globin. *$P < 0.05$, χ^2 test. Embryos were injected with the same sgRNA-Cas9 combinations described in **c** (targeting *dicer1*). (**f**) Fish injected with sgRNA (*dicer1*) and Cas9-nanos or Cas9–β-globin at 4 months of age. The box-and-whisker plot shows the size distribution; *n* denotes the number of embryos analyzed. Boxes span first to last quartiles, and whiskers represent minimum and maximum values. ****$P < 0.0001$, two-tailed Student's *t*-test. (**g**) Pictures of 32-hpf MZ *dicer1* embryos derived from F0 females injected with Cas9-nanos and sgRNA (*dicer1*) outcrossed to dicer1$^{-/-}$ mutant males generated by germline transplantation[28]. Scale bar, 0.5 mm.

targeted gene. To this end, we localized Cas9 expression to the germ line by fusing the *cas9* open reading frame (ORF) to the 3′ UTR of *nanos1* (refs. 22,23) (**Fig. 5b**). To test this method, we mutagenized *dicer1*, *ntla* and *s1pr2*, whose zygotic loss of function impairs larval growth, notochord development and heart development, respectively[24–27]. We compared the activity of Cas9-nanos 3′ UTR (Cas9-nanos) and Cas9–β-globin 3′ UTR (Cas9–β-globin) coinjected into one-cell-stage embryos with sgRNAs to target *ntla* and *dicer1* (**Fig. 5c–f** and **Supplementary Fig. 8a**). The majority of embryos showed the expected phenotype when injected with Cas9–β-globin but not when injected with Cas9-nanos, suggesting that Cas9-nanos reduces the rate of somatic mutations and embryonic lethality. For example, when the *dicer1* locus was targeted, the viability and size of fish injected with Cas9–β-globin were dramatically reduced, and no females were fertile (**Fig. 5e,f**). In contrast, Cas9-nanos–injected fish presented normal growth and homozygous mutant germ cells in 50% of the females, resulting in the generation of maternal zygotic (MZ) *dicer* mutant embryos[28] (**Fig. 5g**). Quantification of MZ *dicer* mutant embryos showed that 12% of the germ line was homozygous *dicer* mutant. Similarly, *ntla* Cas9-nanos–injected fish also had a high rate of germ line transmission for *ntla* mutations (in 50% of the chromosomes) (**Supplementary Fig. 8b**). Together, these results show that germ line targeting of Cas9 can generate MZ and zygotic mutants for genes whose function is required during embryonic and larval development. Coupled with the increased target-site repertoire and optimized sgRNA-sequence design rules, these findings provide a valuable characterization of the elements that influence the activity of the CRISPR-Cas9 system, with the potential to increase both the efficiency and the applicability of CRISPR-Cas9–mediated mutagenesis.

DISCUSSION

Our study provides three insights into the factors that determine the mutagenic activity of the CRISPR-Cas9 system *in vivo*. First, guanine-rich and adenine-depleted sgRNAs are more stable and, correspondingly, more mutagenic than other sgRNAs. We also observed the formation of G-quadruplexes in guanine-rich sgRNAs. Interestingly, G-quadruplexes could also form on the complementary DNA strand at the targeting site. The presence of this structure in double-stranded DNA stabilizes R-loops[29], which are formed in CRISPR-Cas9–DNA interaction complexes[14]. This might also stabilize the sgRNA-Cas9 complex to its cytosine-rich target site, consequently increasing the activity.

Second, we uncovered sequence specificity of Cas9-sgRNA targeting, which was integrated into a predictive model called CRISPRscan. CRISPRscan scores show a stronger correlation with sgRNA activity than offered by previous approaches[11,13], not only in zebrafish but also in *X. tropicalis* (**Supplementary Fig. 9a–d**). Although folding energy is correlated with the stability of the sgRNA, introducing this feature in CRISPRscan did not

improve the predictions for 35 experimentally validated sgRNAs. Furthermore, criteria commonly used to select sgRNA sites, such as an sgRNA GC content between 40% and 80%[12,30], were outclassed by CRISPRscan (**Supplementary Fig. 9e**). Finally, analysis of 132 canonical sgRNAs[31] recently published by the Burgess laboratory showed that CRISPRscan scores were significantly higher ($P = 2.7 \times 10^{-3}$, Mann-Whitney *U*-test; **Supplementary Fig. 9f**) for sgRNAs with higher germ line transmission rates. Thus CRISPRscan provides a valuable resource for predicting the most efficient sgRNA(s) and will facilitate direct functional screenings *in vivo*[32].

Finally, we propose that truncated[20,21] and 5′ mismatch–containing sgRNAs are efficient alternatives to canonical sgRNAs *in vivo*. By increasing the number of available target sites by up to eightfold in the zebrafish genome, we provided more precise targeting of short regions in the genome, such as small ORFs, or specific DNA and RNA functional elements, such as transcription factor–binding sites or microRNA target sites. Furthermore, using Cas9-nanos 3′ UTR fusion allows the generation of MZ mutants in F1, a process that otherwise requires laborious germ line replacement[33]. This method will serve as an entry point for characterizing complete loss-of-function phenotypes during embryogenesis.

Together, this resource provides an accessible framework for designing the most efficient sgRNAs for *in vivo* targeting, particularly in zebrafish, as well as insights into the determinants that mediate sgRNA efficiency.

Online CRISPRscan predictions are available at http://crisprscan.org.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank E. Fleming and H. Codore for technical help; D. Cifuentes, A. Bazzini and M. Lee for discussions; all the members of the Giraldez laboratory for intellectual and technical support; and S. Lau, M. Lee and M. Fernandez-Fuertes for cloning of nanos 3′ UTR, help with MNase analysis and help with pictures of the adult fish, respectively. We thank C. Takacs, M. Lee and K. Divito for manuscript editing. Supported by the Swiss National Science Foundation (grant P2GEP3_148600 to C.E.V.), Programa de Movilidad en Áreas de Investigación priorizadas por la Consejería de Igualdad, Salud y Políticas Sociales de la Junta de Andalucía (M.A.M.-M.), the Fonds de Recherche du Québec (grant 29818 to J.-D.B.) and the US National Institutes of Health (grants R21 HD073768, R01 GM103789, R01 GM102251, R01 GM101108 and GM081602 to A.J.G. and grant R01 HD081379 to E.K.M. and M.K.K.). M.K.K. is supported by the Edward Mallinckrodt Jr. Foundation.

AUTHOR CONTRIBUTIONS

M.A.M.-M., C.E.V. and A.J.G. designed the project, performed experiments and data analysis and wrote the manuscript. J.-D.B. created the sgRNA libraries, performed G-quadruplex experiments and helped write part of the manuscript. J.P.F. performed F0 phenotype analysis and *Xenopus* phenotype analysis with M.A.M.-M. E.K.M. carried out *Xenopus* injections and phenotype analysis. M.K.K. provided reagents and materials.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Bogdanove, A.J. & Voytas, D.F. TAL effectors: customizable proteins for DNA targeting. *Science* **333**, 1843–1846 (2011).
2. Cathomen, T. & Joung, J.K. Zinc-finger nucleases: the next generation emerges. *Mol. Ther.* **16**, 1200–1207 (2008).
3. Hsu, P.D., Lander, E.S. & Zhang, F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157**, 1262–1278 (2014).
4. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
5. Jinek, M. *et al.* RNA-programmed genome editing in human cells. *eLife* **2**, e00471 (2013).
6. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
7. Bassett, A.R. & Liu, J.L. CRISPR/Cas9 and genome editing in *Drosophila*. *J. Genet. Genomics* **41**, 7–19 (2014).
8. Friedland, A.E. *et al.* Heritable genome editing in *C. elegans* via a CRISPR-Cas9 system. *Nat. Methods* **10**, 741–743 (2013).
9. Hwang, W.Y. *et al.* Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nat. Biotechnol.* **31**, 227–229 (2013).
10. Wang, H. *et al.* One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153**, 910–918 (2013).
11. Doench, J.G. *et al.* Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.* **32**, 1262–1267 (2014).
12. Wang, T., Wei, J.J., Sabatini, D.M. & Lander, E.S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80–84 (2014).
13. Gagnon, J.A. *et al.* Efficient mutagenesis by Cas9 protein-mediated oligonucleotide insertion and large-scale assessment of single-guide RNAs. *PLoS One* **9**, e98186 (2014).
14. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
15. Huppert, J.L. Four-stranded nucleic acids: structure, function and targeting of G-quadruplexes. *Chem. Soc. Rev.* **37**, 1375–1384 (2008).
16. Beaudoin, J.D. & Perreault, J.P. Exploring mRNA 3′-UTR G-quadruplexes: evidence of roles in both alternative polyadenylation and mRNA shortening. *Nucleic Acids Res.* **41**, 5898–5911 (2013).
17. Kuscu, C., Arslan, S., Singh, R., Thorpe, J. & Adli, M. Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nat. Biotechnol.* **32**, 677–683 (2014).
18. Farboud, B. & Meyer, B.J. Dramatic enhancement of genome editing by CRISPR/Cas9 through improved guide RNA design. *Genetics* **199**, 959–971 (2015).
19. White, R.M. *et al.* Transparent adult zebrafish as a tool for *in vivo* transplantation analysis. *Cell Stem Cell* **2**, 183–189 (2008).
20. Fu, Y., Sander, J.D., Reyon, D., Cascio, V.M. & Joung, J.K. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat. Biotechnol.* **32**, 279–284 (2014).
21. Ren, X. *et al.* Enhanced specificity and efficiency of the CRISPR/Cas9 system with optimized sgRNA parameters in Drosophila. *Cell Rep.* **9**, 1151–1162 (2014).
22. Köprunner, M., Thisse, C., Thisse, B. & Raz, E. A zebrafish nanos-related gene is essential for the development of primordial germ cells. *Genes Dev.* **15**, 2877–2885 (2001).
23. Mishima, Y. *et al.* Differential regulation of germline mRNAs in soma and germ cells by zebrafish miR-430. *Curr. Biol.* **16**, 2135–2142 (2006).
24. Chen, J.N. *et al.* Mutations affecting the cardiovascular system and other internal organs in zebrafish. *Development* **123**, 293–302 (1996).
25. Halpern, M.E., Ho, R.K., Walker, C. & Kimmel, C.B. Induction of muscle pioneers and floor plate is distinguished by the zebrafish no tail mutation. *Cell* **75**, 99–111 (1993).
26. Schulte-Merker, S. *et al.* Expression of zebrafish goosecoid and no tail gene products in wild-type and mutant no tail embryos. *Development* **120**, 843–852 (1994).
27. Wienholds, E., Koudijs, M.J., van Eeden, F.J., Cuppen, E. & Plasterk, R.H. The microRNA-producing enzyme Dicer1 is essential for zebrafish development. *Nat. Genet.* **35**, 217–218 (2003).
28. Giraldez, A.J. *et al.* MicroRNAs regulate brain morphogenesis in zebrafish. *Science* **308**, 833–838 (2005).
29. Duquette, M.L., Handa, P., Vincent, J.A., Taylor, A.F. & Maizels, N. Intracellular transcription of G-rich DNAs induces formation of G-loops, novel structures containing G4 DNA. *Genes Dev.* **18**, 1618–1629 (2004).
30. Montague, T.G., Cruz, J.M., Gagnon, J.A., Church, G.M. & Valen, E. CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. *Nucleic Acids Res.* **42**, W401–W407 (2014).
31. Varshney, G.K. *et al.* High-throughput gene targeting and phenotyping in zebrafish using CRISPR/Cas9. *Genome Res.* **25**, 1030–1042 (2015).
32. Shah, A.N., Davey, C.F., Whitebirch, A.C., Miller, A.C. & Moens, C.B. Rapid reverse genetic screening using CRISPR in zebrafish. *Nat. Methods* **12**, 535–540 (2015).
33. Ciruna, B. *et al.* Production of maternal-zygotic mutant zebrafish by germ-line replacement. *Proc. Natl. Acad. Sci. USA* **99**, 14919–14924 (2002).

ONLINE METHODS

Target site design. Ten sgRNA target sites were designed within an ~1.2-kb locus for each of 128 loci, for a total of 1,280 sgRNAs. Among these, 1,232 were located in exons, with 1,217 targeting coding sequences. Target sites were spaced by 81 nt on average, with a minimum distance of 18 nt (**Supplementary Fig. 1a**), and had a maximum of ten off-targets with one mismatch. Gene annotations from Ensembl 74 (ref. 34) were used.

Generation of sgRNA and *cas9* mRNA. The sgRNA DNA template was generated by fill-in PCR (**Supplementary Fig. 1e**,**g**). Briefly, a 52-nt oligo (sgRNA primer) containing the T7 promoter (oligo #1; **Supplementary Table 1**), the 20 nt of the specific sgRNA DNA-binding sequence and a constant 15-nt tail for annealing was used in combination with an 80-nt reverse oligo to add the sgRNA invariable 3′ end (tail primer). A 117-bp PCR product was generated according to the following protocol: 3 min at 95 °C; 30 cycles of 30 s at 95 °C, 30 s at 45 °C and 30 s at 72 °C; and a final step at 72 °C for 7 min. PCR products were purified using Qiaquick (Qiagen) columns, and approximately 120–150 ng of DNA were used as a template for a T7 *in vitro* transcription (IVT) reaction (AmpliScribe-T7-Flash transcription kit, Epicentre) (**Supplementary Fig. 1e**). *In vitro* transcribed sgRNAs were treated with DNase and precipitated with sodium acetate and ethanol. Alternative sgRNAs were generated similarly using shorter (50 or 51 nt) and longer (53 or 54 nt) sgRNA primers, with 18–22 nt complementary to the target. sgRNAs beginning with GA sequences contained the SP6 promoter (oligo #2; **Supplementary Table 1**) instead of the T7 promoter. A MAXIscript SP6 transcription kit (Life Technologies) was used for SP6-based IVT reactions.

Zebrafish codon-optimized protein from pT3TS-nCas9n (2656) (ref. 35) was used in all experiments except for the pulldown, where Flag-Cas9 was used. An N-terminal 3×Flag-tag was cloned in pT3TS-nCas9n in the *NcoI* site. The resulting pT3TS-Flag-nCas9n (2722) plasmid was identical to one used previously for a similar experiment in cell lines[4,12]. For the cas9-nanos 3′ UTR construct, the *nanos* 3′ UTR and SV40 late polyA signal were PCR amplified from plasmid pCS2+GFP-nanos 3′ UTR[22,23] using two oligos (oligos #3 and #4; **Supplementary Table 1**). The resulting PCR product was then digested in 3′ with *NotI* and ligated into the pCS2-nCas9n[35] plasmid previously digested with *SnaBI* and *NotI*. The final pCS2-nCas9n-nanos 3′ UTR (2662) (Addgene; 62542) construct was confirmed by sequencing. *cas9* mRNA was *in vitro* transcribed from DNA linearized by either *NotI* (pCS2-nCas9n-nanos 3′ UTR) or *XbaI* (pT3TS-nCas9n and pT3TS-FLAG-nCas9n) using the mMachine SP6 or T3 kit (Ambion), respectively. *In vitro* transcribed mRNAs were treated with DNase and purified using the RNeasy Mini kit (Qiagen).

RNA injections, large-scale experiments and plasmids. For large-scale experiments, 1,280 sgRNAs (**Supplementary Table 1**) were injected in 16 different cocktails. Independent injections containing 240 pg of 80 sgRNAs targeting eight genes and 300 pg of *cas9* mRNA were carried out at the one-cell stage. Five embryos per injection (80 in total) were collected at 1.25 hpf for sgRNA input. Seven embryos per injection were collected at 6 and 9 hpf (102 in total) for the Cas9 loading experiment (described below). Twenty embryos per injection were collected at 9 hpf, and DNA was extracted according to the HotShot protocol[36], with minor modifications. Briefly, an ~1.2-kb PCR product was obtained for each of the 128 loci (**Supplementary Table 1**) using the following protocol: 3 min at 95 °C; 35 cycles of 30 s at 95 °C, 30 s at 60 °C and 2 min at 72 °C; and a final step at 72 °C for 7 min. We collected and analyzed 40 noninjected embryos to determine possible polymorphisms. PCR products were visualized and quantified on agarose gel (Adobe Photoshop). Next, similar amounts of PCR products per gene were pooled and purified (QIAquick PCR purification, Qiagen). We sheared purified amplicons to obtain 150-pb DNA products, which we used to generate DNA libraries (described below). The same approach was used for the alternative sgRNA experiment, but with 640 sgRNAs in eight different injections (**Supplementary Table 1**).

To compare the efficiency of Cas9-nanos with that of Cas9–β-globin in phenotype-analysis experiments and in the CRISPRscan independent validation experiments, we injected 100 pg of *cas9* mRNA and 20 or 10 pg of each sgRNA at the one-cell stage. Phenotypes were analyzed and quantified between 24 and 48 hpf.

Flag-Cas9 immunoprecipitation. Flag-Cas9 immunoprecipitation was performed as previously described[37], with some modifications. Briefly, 102 Flag-Cas9–injected embryos were collected at either 6 or 9 hpf and flash-frozen in liquid nitrogen. Embryos were lysed in 1 ml of NET-2 buffer (100 mM Tris-HCl, pH 7.5, 150 mM NaCl and 0.05% NP-40) supplemented with protease and RNase inhibitor (Roche). One-fiftieth (25 µl) of the lysate was collected as an input control, and the remaining fraction was added to 100 µl of Flag-M2 magnetic beads (Sigma) previously washed three times with NET-2 buffer. Samples were incubated at 4 °C for 2 h with orbital shaking. After incubation, another 25-µl aliquot was collected from the supernatant as a supernatant control. Beads were then washed four times with 1 ml of NET-2 buffer at 4 °C. A final 25-µl aliquot was collected during the last wash as a pulled-down control. Next, 500 µl of TRIzol (Life Technologies) was added to the beads, and RNA was purified and used as starting material for sgRNA cloning (described below). The same protocol was applied to 102 noninjected embryos as a negative control. Finally, the input, supernatant and pulled-down controls were subjected to SDS-PAGE and analyzed by western blot. Mouse monoclonal anti–Flag-M2 (Sigma-Aldrich, F1804, 1:2,000) and rabbit polyclonal anti-gamma (Abcam, ab11317, 1:20,000) were used according to the manufacturer's instructions.

Cloning of the sgRNAs. Total RNA was isolated from embryos injected with sgRNAs using TRIzol reagent (Life Technologies). Purified total RNA was then subjected to reverse transcription using the SuperScript III First Strand kit (Invitrogen), according to the manufacturer's protocol, and a primer containing a 3′ sequence complementary to the constant part of the sgRNAs and a 5′ sequence corresponding to part of the Illumina TruSeq Adaptor (oligo #5; **Supplementary Table 1**). The resulting first-strand cDNAs were purified with the Agencourt AMPure XP system (Beckman Coulter) according to the manufacturer's protocol. cDNAs were dissolved in 10 µL of water, and an ssDNA linker was added at their 3′ ends (5′-/5Phos/linker/3InvdT/-3′, where /5Phos/ is a 5′ phosphate and /3InvdT/ is an inverted deoxythymidine; the linker is presented in **Supplementary Table 1** (oligo #6))

using CircLigase ssDNA Ligase (Epicentre), with slight modifications to the manufacturer's protocol. In brief, reagents were added to 3 µL of dissolved cDNA samples to reach the following final concentrations in a total volume of 10 µL: 1× CircLigase buffer, 0.05 mM ATP, 2.5 mM $MnCl_2$, 10% PEG 6000, 1 M betaine, 5 µM ssDNA linker and 50 U CircLigase enzyme. The ligation mixture was then incubated at 60 °C for 2 h, 68 °C for 1 h and 80 °C for 10 min to deactivate the CircLigase. The volume was increased by the addition of 10 µL of water, and ligated products were purified with the Agencourt AMPure XP system (Beckman Coulter) according to the manufacturer's protocol. PCR amplification was performed on the ligated product using Illumina primers (Illumina small RNA forward primer and Illumina TruSeq reverse index in **Supplementary Table 1** (oligos #7 and #8); indexes 6, 7, 12–16, 19, 23, 25 and 27 were used in this study). PCR products were purified on an 8% native polyacrylamide gel, and bands corresponding to the sgRNA final library size (179 nt) were extracted. DNA was eluted, ethanol precipitated, dissolved in water and sent for sequencing.

For input samples, *in vitro* transcribed sgRNA cocktails used for injection were pooled together and diluted 1/100, and 1 µL of this dilution was combined with total RNA isolated from uninjected embryos before reverse transcription. Alternative sgRNA samples and RNA samples isolated from the Cas9 pulldown experiment were treated as described above, with a few modifications. Briefly, 50 U of RNase I was added during the RNase H treatment before AMPure XP purification. Finally, the AMPure XP purification prior to the PCR amplification was replaced by a 10% urea-denaturing PAGE purification step. Bands corresponding to the ligated product size (114 nt) were cut. DNA was then eluted, ethanol precipitated, dissolved in water and PCR amplified as described above.

sgRNA labeling and in-line probing. To produce 5′-end-labeled sgRNAs, we dephosphorylated purified sgRNA transcripts by adding 1 U of antartic phosphatase (New England BioLabs) to 50 pmol of sgRNA in a final volume of 10 µL containing 50 mM Bis-propane, pH 6.0, 1 mM $MgCl_2$, 0.1 mM $ZnCl_2$ and RNase OUT (20 U; Invitrogen). The mixture was incubated for 30 min at 37 °C, and then the enzyme was inactivated by incubation for 5 min at 65 °C. Dephosphorylated sgRNAs (5 pmol) were 5′-end radiolabeled with 3 U of T4 polynucleotide kinase (New England BioLabs) for 1 h at 37 °C in the presence of 3.2 pmol of [α-^{32}P]ATP (6,000 Ci/mmol; PerkinElmer). The reaction was stopped by the addition of formamide dye buffer (95% formamide, 10 mM EDTA, 0.025% bromophenol blue and 0.025% xylene cyanol), and the radiolabeled sgRNA was purified by 8% PAGE. 5′-end-labeled sgRNAs were visualized by autoradiography. The bands corresponding to the correct sizes were excised from the gel, gel slices were shredded, and the transcripts were eluted for 10 min at 37 °C in 300 µL of water. The labeled sgRNAs were then ethanol precipitated, washed and dissolved in water.

In-line probing was performed to test for G-quadruplex formation by sgRNAs as described previously[38]. This technique and the conditions used allowed the study of the RNA structure in conditions that either did (KCl) or did not (LiCl) support G-quadruplex formation. Because G-quadruplexes are the only potassium-dependent structure at the concentrations used, a sequence can be considered as having formed such a structure if a difference in the band patterning is observed[38]. Briefly, 5′-end-labeled sgRNAs (50,000 cpm) corresponding to a trace amount of RNA (<1 nM) were heated at 70 °C for 5 min and then slow-cooled to room temperature over 1 h in buffer containing 50 mM Tris-HCl, pH 7.5, in the presence of either 100 mM LiCl or 100 mM KCl in a final volume of 10 µL. After incubation, the final volume of each sample was adjusted to 100 µL to reach final concentrations of 50 mM Tris-HCl, pH 7.5, 20 mM $MgCl_2$ and either 100 mM LiCl or 100 mM KCl. The mixtures were incubated for 40 h at room temperature and ethanol precipitated, and the RNAs were dissolved in formamide dye loading buffer (95% formamide, 10 mM EDTA, 0.025% bromophenol blue and 0.025% xylene cyanol). For alkaline hydrolysis, 50,000 cpm of 5′-end-labeled sgRNAs (<1 nM) were dissolved in 5 µL water, 1 µL of 1 M NaOH was added, and the reactions were incubated for 30 s at room temperature before being quenched by the addition of 3 µL of 1 M Tris-HCl, pH 7.5. The RNAs were then ethanol precipitated and dissolved in formamide dye loading buffer. RNase T1 ladders were prepared using 50,000 cpm of 5′-end-labeled RNA (<1 nM) dissolved in buffer containing 20 mM Tris-HCl, pH 7.5, 10 mM $MgCl_2$ and 100 mM LiCl. The mixtures were incubated for 1.5 min at 37 °C in the presence of 3 U of RNase T1 (Fisher Scientific) and were stopped by ethanol precipitation before being dissolved in formamide dye loading buffer. The radioactivity of the in-line probing samples and both ladders was calculated, and equal amounts in terms of the counts per minute of all conditions and ladders of each sgRNA were fractionated on denaturing (8 M urea) 10% polyacrylamide gels. The resulting gels were subsequently dried and visualized by exposure to a phosphor screen. The band intensities for each condition were calculated using the Semi-Automated Footprinting Analysis (SAFA) software. sgRNAs showing a ratio of band intensities (KCl/LiCl) greater than 2 for a nucleotide in or flanking the guanine-rich sequence were identified as positively folding into a G-quadruplex structure. A previous study showed that a threshold of 2 was associated with guanine-rich sequences folding into active G-quadruplexes *in cellulo*[39].

Library preparation and sequencing of genomic loci. We determined amplicon quality and concentration by estimating the A260/A280 and A260/A230 ratios by nanodrop. We confirmed amplicon integrity and size by running an Agilent Bioanalyzer gel. 100 ng of PCR amplicons (1–2 kb) were sheared to 120–150 bp with a Covaris E210 instrument. After shearing, solid-phase reversible immobilization bead cleanup was performed using an Ampure XP SPRI (Beckman Coulter Genomics). Shearing quality control was then performed on sheared products using the DNA 1000 Bioanalyzer chip to confirm the target size. Sheared amplicons were then end-repaired and A-tailed, and adapters were ligated. Indexed libraries that met appropriate cutoffs were quantified by both quantitative RT-PCR using a commercially available kit (KAPA Biosystems) and insert-size distribution determined with a LabChip GX. Samples with a yield of ≥0.5 ng/µl were used for sequencing. Sample concentrations were normalized to 2 nM and loaded onto Illumina version 3 flow cells at a concentration that yielded 170 million to 200 million passing filter clusters per lane. Samples were then sequenced using 75-bp paired-end reads on an Illumina HiSeq 2000 according to Illumina protocols. The 6-bp index was read during an additional sequencing read

doi:10.1038/nmeth.3543

that automatically followed the completion of read 1. A positive control (prepared bacteriophage Phi X library) provided by Illumina was spiked into every lane at a concentration of 0.3% to allow monitoring of sequencing quality in real time. Primary analysis and sample demultiplexing were performed using Illumina's CASAVA 1.8.2 software suite.

Raw reads are publicly accessible in the Sequence Read Archive under SRP059430.

Zebrafish maintenance. Wild-type zebrafish embryos were obtained through natural mating of TU-strain zebrafish of mixed ages (5–18 months) for large-scale experiments, except for the Cas9-Flag pulldown, where TU-AB and TLF strains of mixed ages (5–17 months) were used. TU-AB and TLF strains of mixed ages were also used for the other experiments. Selection of mating pairs was random from a pool of 60 males and 60 females allocated for a given day of the month. Fish lines were maintained in accordance with research guidelines of the International Association for Assessment and Accreditation of Laboratory Animal Care, under a protocol approved by the Yale University Institutional Animal Care and Use Committee (IACUC).

Frog husbandry and injections. *X. tropicalis* were housed and cared for in our aquatics facility according to established protocols that were approved by the Yale University IACUC.

Ovulation was induced and eggs were collected according to established protocols[40]. Staging of *Xenopus* tadpoles was done according to the procedure described by Nieuwkoop and Faber[41]. mRNA was injected into one-cell or two-cell embryos as previously described[42], along with dextran mini-Ruby as a tracer. 500 pg of *cas9* mRNA and 400 pg of each sgRNA were injected into one-cell-stage embryos.

Image acquisition. Embryos were analyzed using a Zeiss Axioimager M1 and Discovery microscopes and photographed with a Zeiss Axiocam digital camera. Images were processed with Zeiss AxioVision 3.0.6. Adult fish were photographed with a Panasonic Lumix DMC-FZ18.

Mapping of wild-type and mutant reads. To filter the reads that potentially were mutated by CRISPR-Cas9, we mapped all reads to the zebrafish genome *Zv9* using Bowtie2 2.2 (ref. 43), not allowing any insertions or deletions and increasing mutation costs with the following options: -end-to-end,-rdg 1000,100,-rfg 1000,100,-mp 12,4. Reads that aligned to the loci were counted as 'wild-type' reads to be used as a reference for sgRNA activity (discussed below). We filtered PCR oligos from the unmapped pairs (one or both unmapped mates) by trimming the oligos. Reads were first mapped to the oligos with Bowtie2 with the following options: -local, -L 4, -a,-score-min G,8,4,-rfg 10,5,-rdg 10,5,-reorder. Then unmapped reads were kept together with trimmed reads if fewer than two oligos aligned to the read, the oligo aligned within the first 4 nt of the read, and the trimmed read was longer than 30 nt. Reads were aligned to the sequence of the 128 loci using gmap 2014.10.22 (ref. 44). The alignments with gmap allowed the detection of distinct mutations on the same read caused by different sgRNAs. Reads aligned as concordant pairs on one of the loci were kept for the rest of the analysis. For an insertion or deletion event to be called, a 15-nt match was required on each side of the called mutation in the read alignment.

Characterization of single and double sgRNA mutations. A region centered on the known Cas9 cleavage site (6 nt upstream of the PAM[14]), extended by 15 nt upstream and downstream, was used to attribute mutations to a specific sgRNA target site. A mutation starting or ending in one of the 1,280 regions was attributed to that sgRNA and counted as an sgRNA-mediated mutation. A mutation starting and ending in two different regions was counted as a 'two-site mutation'. A mutation that overlapped a single region was counted as a 'single-site mutation'. Otherwise, it was counted simply as a mutation (overlapping more than one region). To analyze the length and frame distribution of the mutations, we discarded the top 1% of outliers.

Detection of polymorphisms. Reads of the noninjected controls were mapped similarly to those for the sgRNA-injection experiment. Mapped reads were then piled up with SAMtools[45], and all variants were detected using BCFtools with the following options: -mv -P1e-3. A minimum quality of 5, a 150-read coverage and a 50% frequency were required for a polymorphism to be called. Target sites overlapping with at least one polymorphism were discarded.

Measuring sgRNA input. Reads were mapped to the sgRNA sequences using Bowtie2 configured to perform local alignments for automatic adaptor trimming with increased gap and mismatch costs with the following options: -local, -L 4, -a,-score-min G,8,4,-rfg 10,5,-rdg 10,5. Mapped reads (only primary alignment as defined in the SAM format specifications to ensure that in the rare cases where a sequencing read aligned partially to multiple sgRNAs, only the correct alignment was selected) for each sgRNA were then counted. As sgRNAs were injected by a cocktail of 80, counts were normalized by the total number of reads in each cocktail.

Computing sgRNA folding energy. RNAfold from the ViennaRNA package[46] was used with the -partfunc parameter to compute the EFE.

Computing sgRNA activity. To obtain robust sgRNA raw activity, we required a minimum of 1,000 reads overlapping each target-site region, as well as a minimum of 150 reads in the noninjected controls to ensure that any polymorphism could be detected. The raw activity was computed as the percentage of mutated reads over wild-type reads and mutated reads overlapping each sgRNA region as described above. To obtain robust sgRNA normalized activity, we removed the least abundant 5% of sgRNAs (on the basis of normalized counts at 0 hpf or 1.25 hpf). This step also reduced the noise level of sgRNA abundances. The normalized sgRNA activity was computed as the raw activity divided by the $\log_2$ of the sgRNA normalized count. We then computed a rank percentile. The sgRNA with the highest normalized activity received a rank of 1.

sgRNA-activity regression model. To build the linear regression model, we included 684 features corresponding to the sequence

of the sgRNA, the PAM and six upstream and downstream nucleotides. These 684 features consisted of the following: 140 features representing mononucleotides (4 base identities × 35 nucleotide positions (6 nt upstream context + 20 nt sgRNA + 3 nt PAM + 6 nt downstream context)) and 544 dinucleotide features (16 possible dinucleotides × 34 positions). Each sgRNA plus context (35 nt total) was represented as a binary vector of length 684 encoding the presence (value of 1) or absence (value of 0) for each feature. A randomized logistic regression[47,48] with 500-fold resampling, L1 penalty and 0.3 regularization strength was used to select 91 features that were determinant to classify the top 20% of the most efficient sgRNAs. A linear regression was then fitted on the 91 features and the ranked normalized sgRNA activities. Cross-validation was performed using the ShuffleSplit method of scikit-learn with 200 iterations.

The canonical model was applied directly on mismatch-containing alternative sgRNAs; correlations between experimental rank and CRISPRscan-predicted score for Gg18 (r = 0.57, P = 0.004) and gG18 (r = 0.26, P = 0.06) alternatives were observed. To apply the canonical scoring model to shorter sgRNAs, we explored two strategies to account for missing nucleotide positions encoded in the GG18 model. An empirical approach evaluating performance was employed wherein (a) index positions were omitted from the encoding of shorter sgRNAs and thus not scored, or (b) index positions were assigned best-approximation base identities, defined by the next nucleotide downstream (i.e., base X could contribute to nucleotide position i or i + 1 in the canonical encoding) (**Supplementary Fig. 6i**). Because the biophysical significance of base preferences at each nucleotide position is unclear, these options did not favor any *a priori* assumptions regarding equivalent nucleotide indices between shorter and canonical-length sgRNAs, and instead allowed for a slightly 'fuzzy' encoding.

MNase. MNase sequencing data were obtained from GEO accession GSE44269 and analyzed similarly as described by Zhang *et al.*[49]. Normalized read coverage was summed over the sgRNA target-site regions.

Cas9 distraction factor. Off-target seeds were the number of 5 + (N + 2) nt or 7 + (N + 2) nt (where N + 2 is the PAM) occurrences found on both strands of the zebrafish genome. For the control, the seeds were on the opposite side of the PAM in the sgRNA target.

Code availability. The mutagenesis analysis pipeline is available upon request.

Bioinformatics libraries. Python custom scripts were used to perform the analysis using the Python libraries Matplotlib (http://matplotlib.org) for plotting, Numpy (http://numpy.org) and Pandas (http://pandas.pydata.org) for data mining, Scipy (http://scipy.org) for statistics, and scikit-learn (http://scikit-learn.org) and Statsmodels (http://statsmodels.sourceforge.net) for machine learning.

34. Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Res.* **43**, D662–D669 (2015).
35. Jao, L.E., Wente, S.R. & Chen, W. Efficient multiplex biallelic zebrafish genome editing using a CRISPR nuclease system. *Proc. Natl. Acad. Sci. USA* **110**, 13904–13909 (2013).
36. Meeker, N.D., Hutchinson, S.A., Ho, L. & Trede, N.S. Method for isolation of PCR-ready genomic DNA from zebrafish tissues. *Biotechniques* **43** 610, 612, 614 (2007).
37. Cifuentes, D. *et al.* A novel miRNA processing pathway independent of Dicer requires Argonaute2 catalytic activity. *Science* **328**, 1694–1698 (2010).
38. Beaudoin, J.D., Jodoin, R. & Perreault, J.P. In-line probing of RNA G-quadruplexes. *Methods* **64**, 79–87 (2013).
39. Beaudoin, J.D. & Perreault, J.P. 5′-UTR G-quadruplex structures acting as translational repressors. *Nucleic Acids Res.* **38**, 7022–7036 (2010).
40. del Viso, F., Bhattacharya, D., Kong, Y., Gilchrist, M.J. & Khokha, M.K. Exon capture and bulk segregant analysis: rapid discovery of causative mutations using high-throughput sequencing. *BMC Genomics* **13**, 649 (2012).
41. Nieuwkoop, P.D. & Faber, J. *Normal Table of Xenopus laevis (Daudin): A Systematical and Chronological Survey of the Development from the Fertilized Egg till the End of Metamorphosis* (Garland, 1994).
42. Khokha, M.K. *et al.* Techniques and probes for the study of *Xenopus tropicalis* development. *Dev. Dyn.* **225**, 499–510 (2002).
43. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
44. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
45. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
46. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
47. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* (Springer, 2001).
48. Meinshausen, N. & Bühlmann, P. Stability selection. *J. R. Stat. Soc. Series B Stat. Methodol.* **72**, 417–473 (2010).
49. Zhang, Y. *et al.* Canonical nucleosome organization at promoters forms during genome activation. *Genome Res.* **24**, 260–266 (2014).

doi:10.1038/nmeth.3543